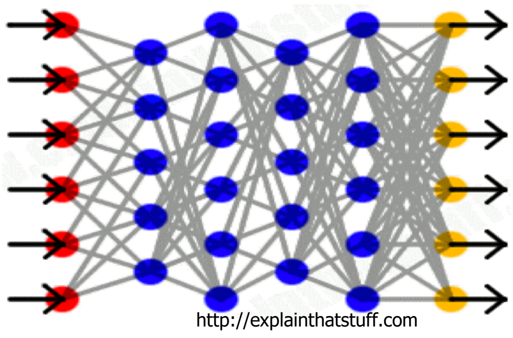
Explainable Artificial Intelligence (XAI)

David Gunning
Information Innovation Office (I2O)
Defense Advanced Research Projects Agency (DARPA)

The Need for Explainable AI

EXPLAINABLE ARTIFICIAL INTELLIGENCE

AI System



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

Transportation

Finance

Security

Legal

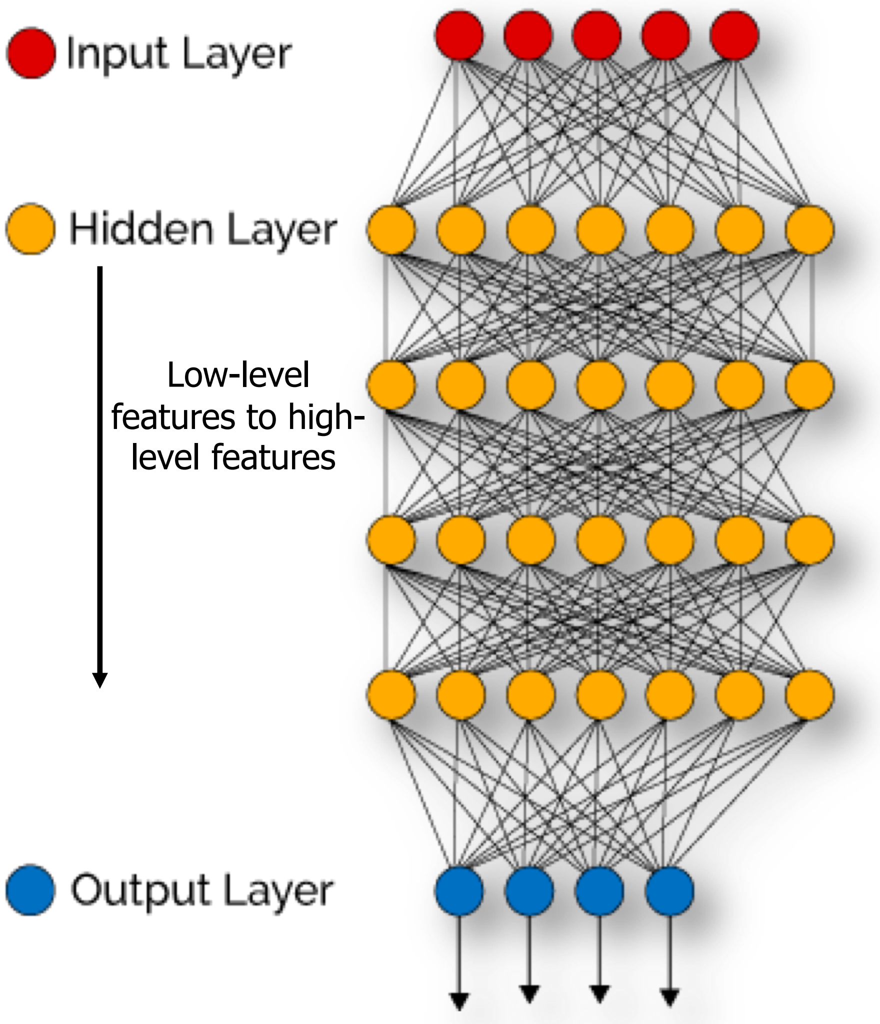
Medicine

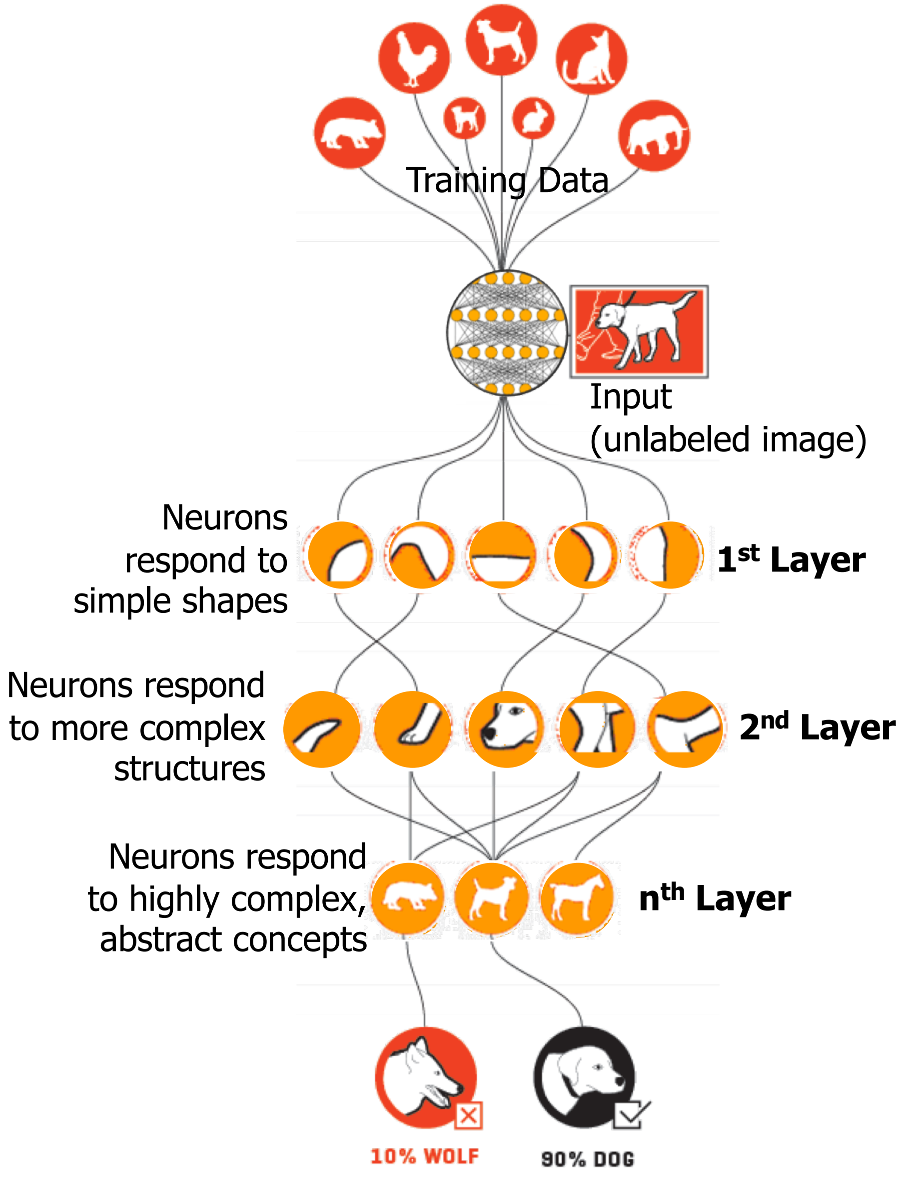
Military

User

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

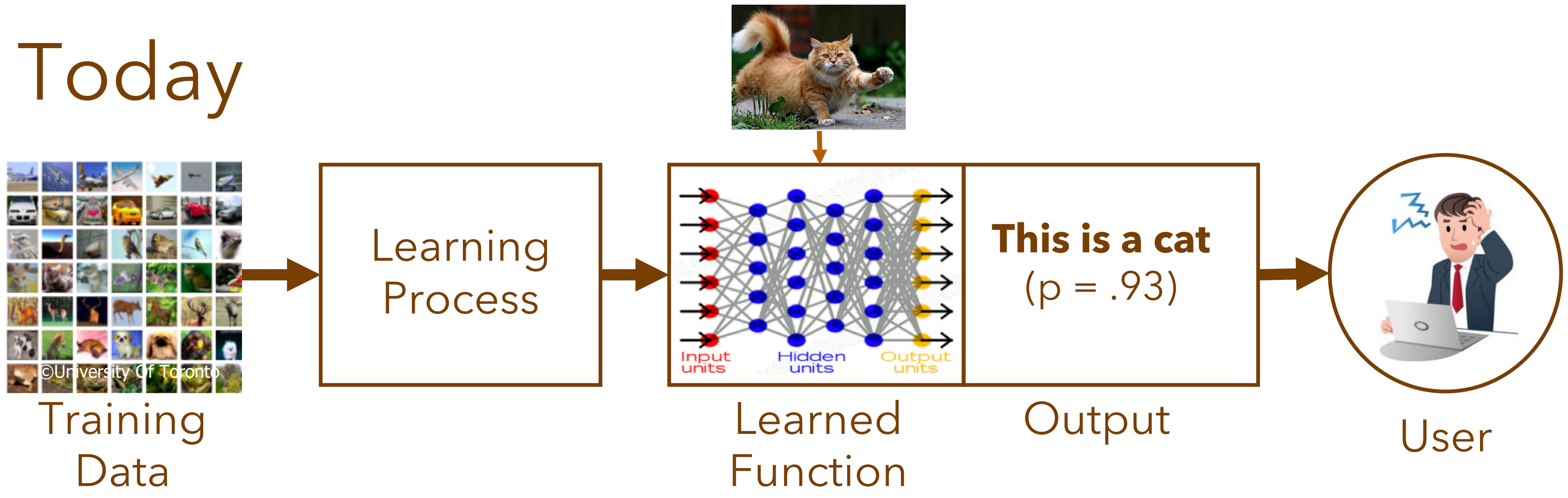
Deep Learning Neural Networks





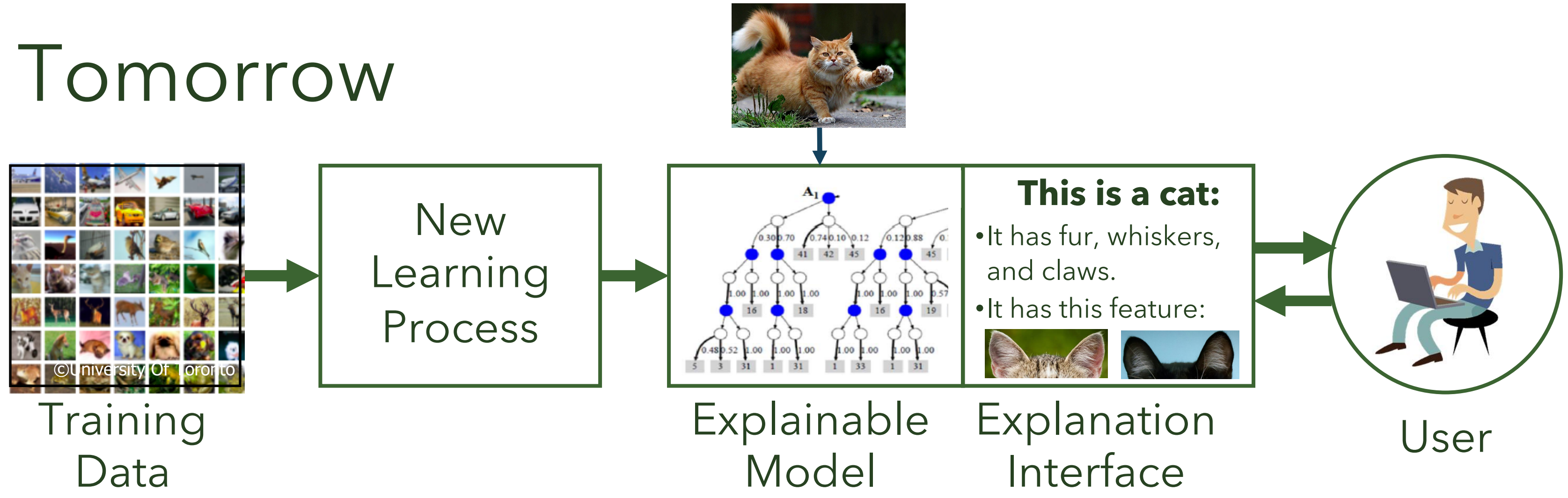
What are we trying to do?

Today



- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

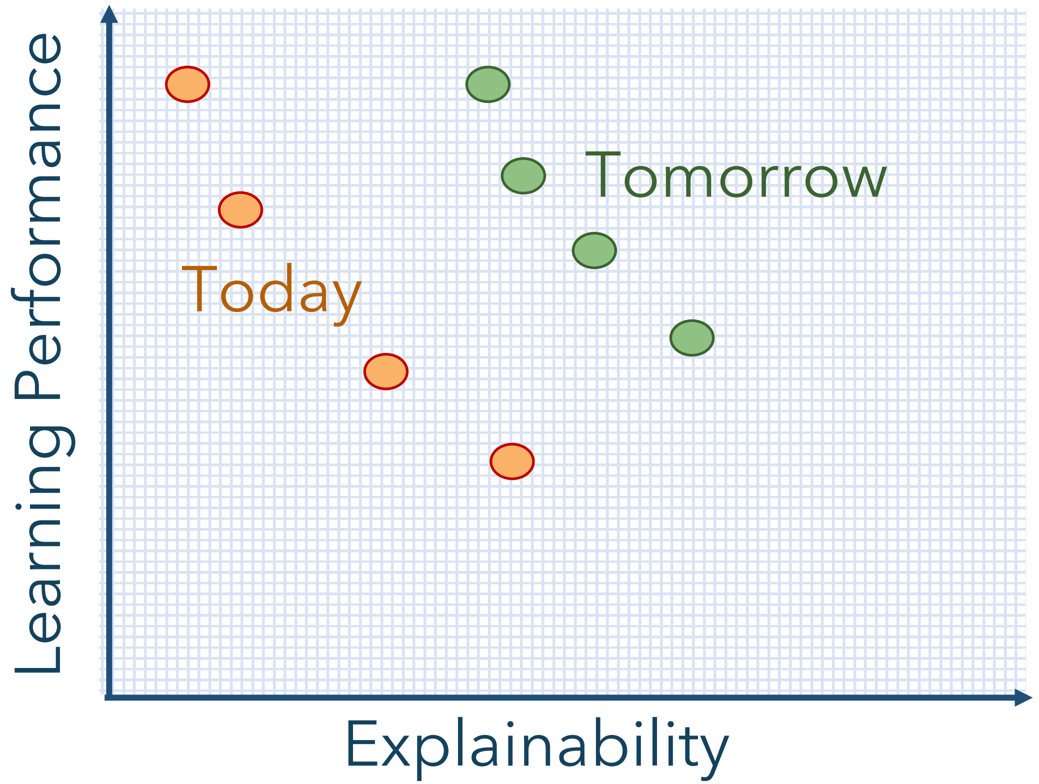
Tomorrow



- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

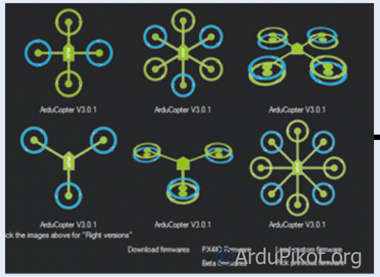
Goal: Performance and Explainability

- XAI will create a suite of machine learning techniques that
  - Produce more explainable models, while maintaining a high level of learning performance
  - Enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems



Challenge Problems

EXPLAINABLE ARTIFICIAL INTELLIGENCE

	Learn a model	**Explain decisions**	**Perform a task**	
Data Analytics Classification Learning Task	 Multimedia Data	Explainable Model → Explanation Interface	Recommend / Explanation 	An analyst is looking for items of interest in massive multimedia data sets
	Classifies items of interest in large data set	Explains why/why not for recommended items	Analyst decides which items to report, pursue	
Autonomy Reinforcement Learning Task	 ArduPilot & SITL Simulation	Explainable Model → Explanation Interface	Actions / Explanation 	An operator is directing autonomous systems to accomplish a series of missions
	Learns decision policies for missions	Explains behavior in an after-action review	Operator decides which future tasks to delegate	

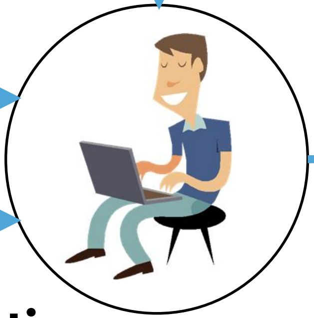
Measuring Explanation Effectiveness

Explanation Framework

Task

Recommendation, Decision or Action

Explainable Model | Explanation Interface



Decision

The user makes a decision based on the explanation

XAI System

The system takes input from the current task and makes a recommendation, decision, or action

Explanation

The system provides an explanation to the user that justifies its recommendation, decision, or action

User Satisfaction

- Clarity of the explanation (user rating)
- Utility of the explanation (user rating)

Mental Model

- Understanding individual decisions
- Understanding the overall model
- Strength/weakness assessment
- 'What will it do' prediction
- 'How do I intervene' prediction

Task Performance

- Does the explanation improve the user's decision, task performance?

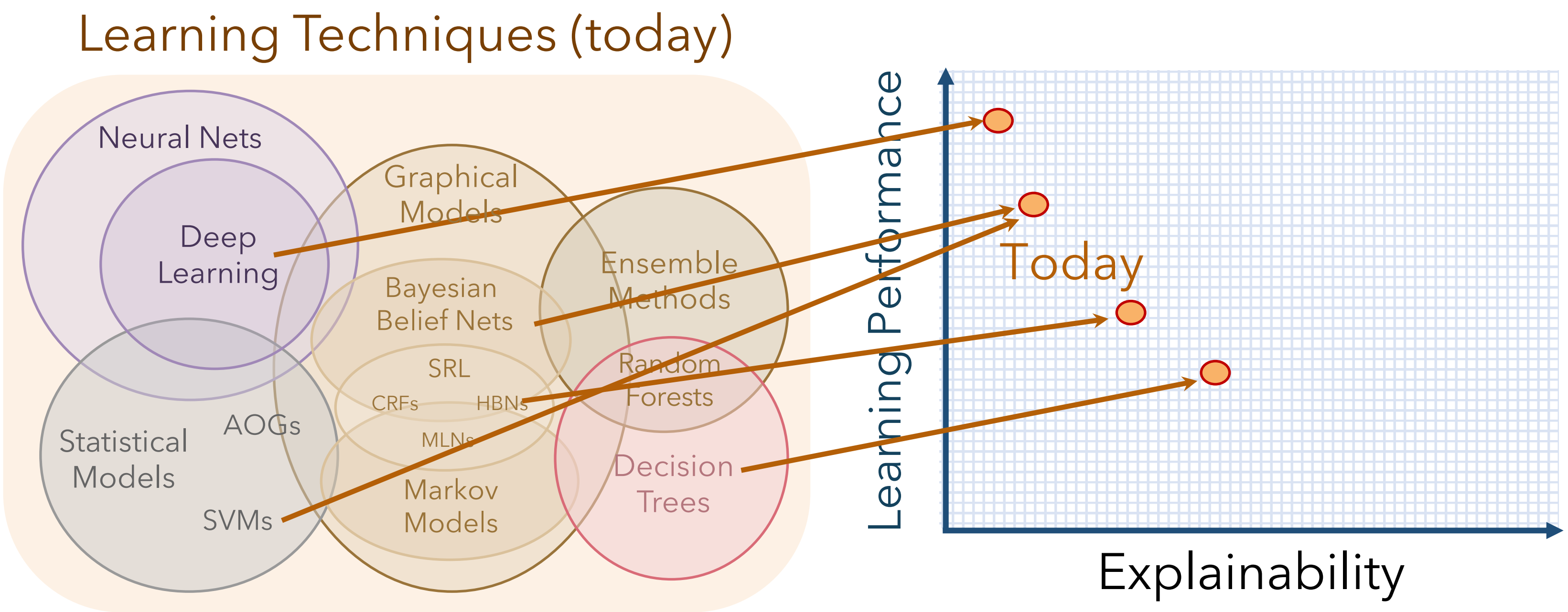
Trust Assessment

- Appropriate future use and trust

Correctability (Extra Credit)

- Identifying errors
- Correcting errors

Developing an Explainable Model



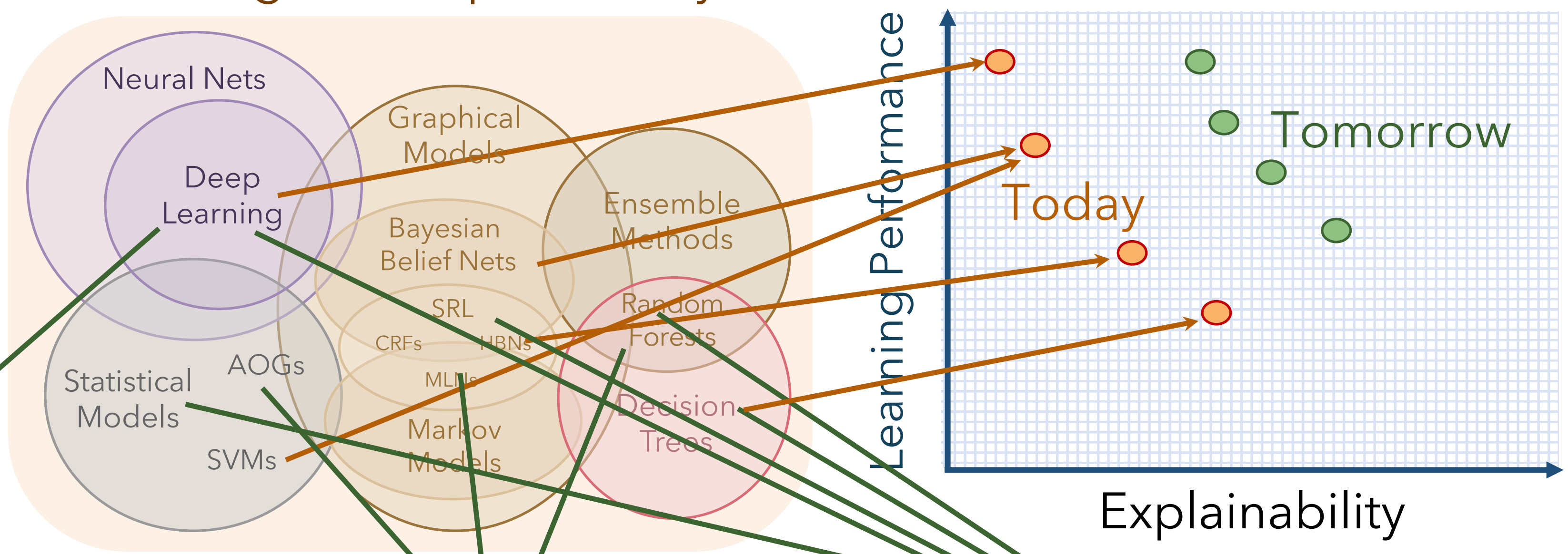
Developing an Explainable Model

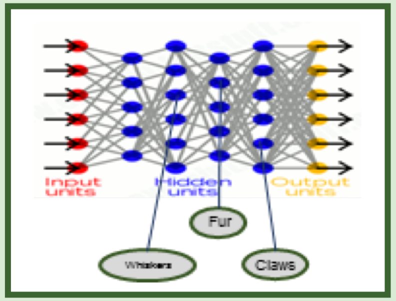
EXPLAINABLE ARTIFICIAL INTELLIGENCE

XAI Goal

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

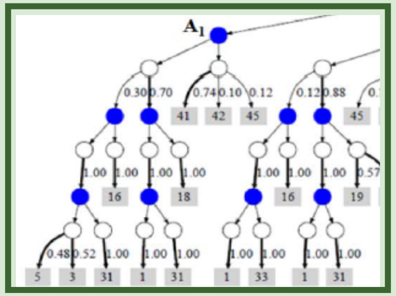
Learning Techniques (today)





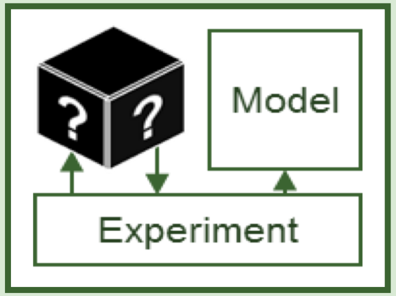
Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

Techniques to learn more structured, interpretable, causal models



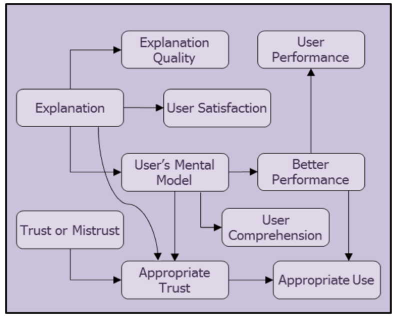
Model Induction

Techniques to infer an explainable model from any model as a black box

XAI Developers and Technical Approaches

New Learning Process → Explainable Model | Explanation Interface ⇄

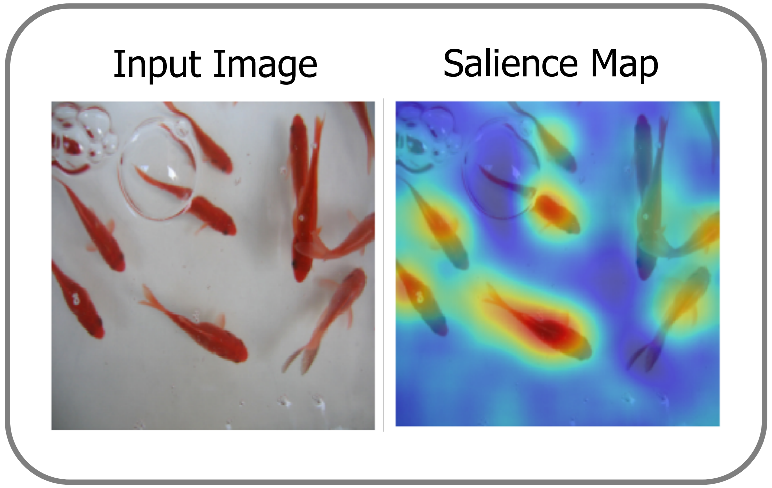
CP	Performer	Explainable Model	Explanation Interface
Both	UC Berkeley	Deep Learning	Reflexive and Rational
	Charles River	Causal Modeling	Narrative Generation
	UCLA	Pattern Theory+	3-level Explanation
Autonomy	Oregon State	Adaptive Programs	Acceptance Testing
	PARC	Cognitive Modeling	Interactive Training
	CMU	Explainable RL (XRL)	XRL Interaction
Analytics	SRI International	Deep Learning	Show and Tell Explanation
	Raytheon BBN	Deep Learning	Argumentation and Pedagogy
	UT Dallas	Probabilistic Logic	Decision Diagrams
	Texas A&M	Mimic Learning	Interactive Visualization
	Rutgers	Model Induction	Bayesian Teaching



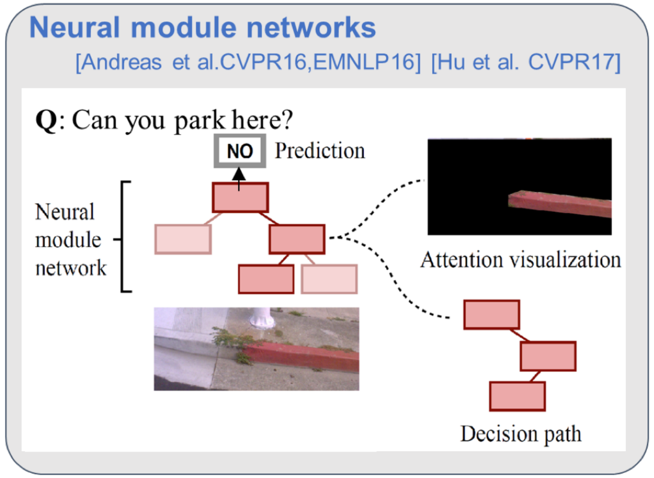
IHMC
Psychological Models of Explanation

Approaches to Deep Explanation

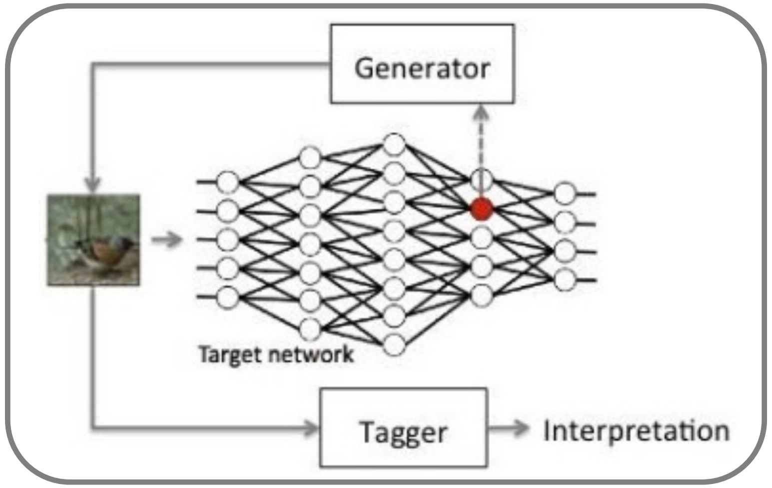
Attention Mechanisms



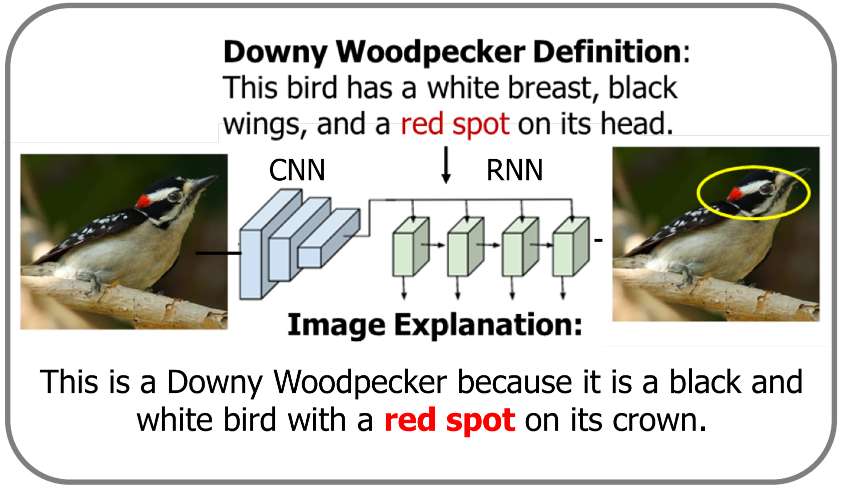
Modular Networks



Feature Identification



Learn to Explain



Deeply Explainable Artificial Intelligence

UC Berkeley, Boston U., U. Amsterdam, Kitware

Explainable Model	Explanation Interface	Challenge Problem

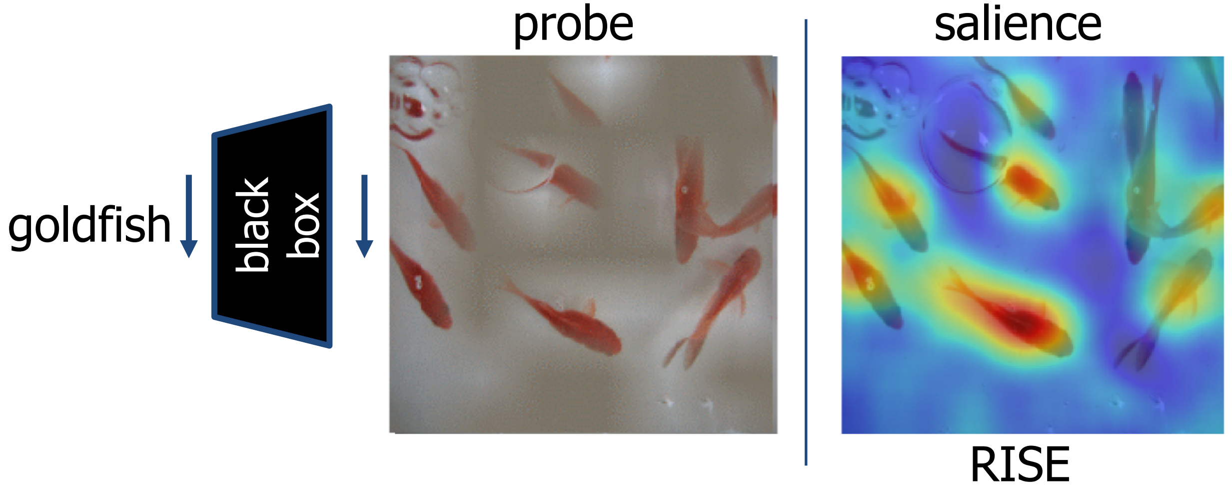
Deep Learning
• Post-hoc explanations by training additional DL models
• Explicit introspective explanations (Neural Module Networks)
• Reinforcement Learning
  • Informative rollouts
  • Explicit modular agent
	Reflexive and Rational
	• Reflexive explanations (arise from the model)
	• Rational explanations (come from reasoning about user's beliefs)
	• Evaluation criteria
	  • Human interpretability
	  • Predictive behavior
	  • Appropriate trust
		Autonomy
		• Vehicle control (BDD-X, CARLA)
		• Strategy games (StarCraft II)
		
		Data Analytics
		• Visual QA and filtering tasks (VQA-X, ACT-X, xView, DiDeMo, etc.)

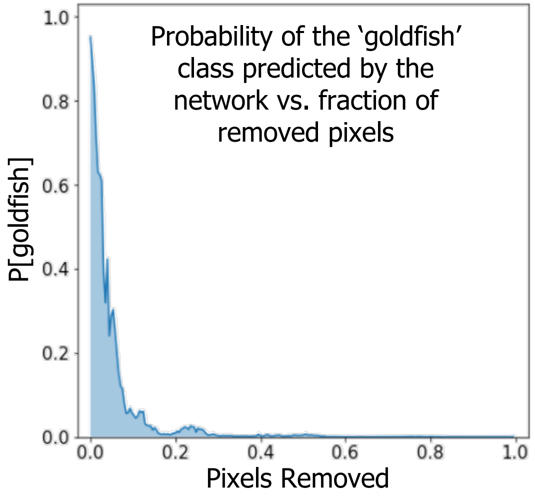
- **PI**: Trevor Darrell (UC Berkeley)

- Pieter Abbeel (UC Berkeley)
- Tom Griffiths (UC Berkeley)
- Kate Saenko (Boston U.)
- Zeynep Akata (U. Amsterdam)
- Dan Klein (UC Berkeley)
- John Canny (UC Berkeley)
- Anca Dragan (UC Berkeley)
- Anthony Hoogs (Kitware)

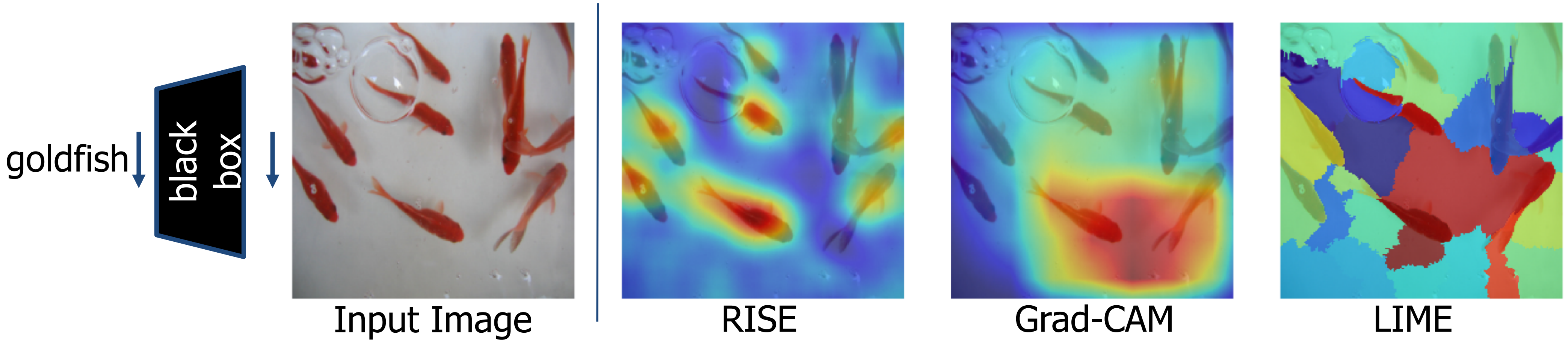
High Fidelity Visual Salience Model

UC Berkeley, Boston U., U. Amsterdam, Kitware





Salience is estimated by probing the black box with randomly masked images



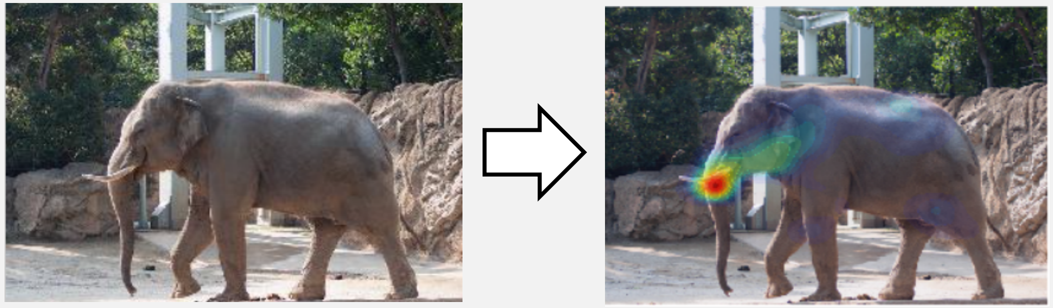
Petsiuk, Das and Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models, 2018

Textual justification with grounding and pragmatics

Given the multi-modal explanation generated by the model, do you think the system will answer correctly?

Question: *Does this elephant have tusks?*

"because there are no bones sticking out from its mouth"



Yes No

Incorrect! The system answered *"no"* when the ground-truth answer is *"yes"*

Question: *Is this a professional sporting event?*

"because the players are wearing official jerseys"

Yes No

Correct! The system answered *"yes"* when the ground-truth answer is *"yes"*

Explanation Effectiveness	Attention for Explanation Used?	Accuracy of Users Judgement
Without explanation (existing SOTA)	No	57.5%
UCB Model on descriptions	Yes	66.5%
UCB Model without attention	No	61.5%
UCB Model	Yes	**70.0%**

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence, 2018

Causal Grounded Driving

acceleration course

Vehicle controller

Predicts vehicle control commands with a visual attention model

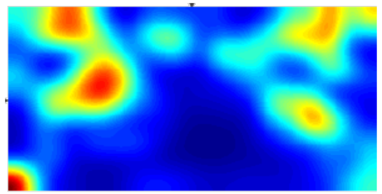
control outputs (steering angle)

Explanation generator

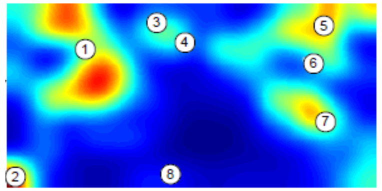
Attention-based video-to-text model to produce textual justifications

descriptions + explanations

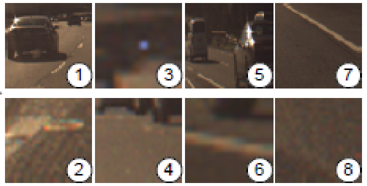
Input image stream



Attention heat map



Clustering analysis



Visual saliency detection and causality check

Kim and Canny, in ICCV, 2017
Kim, Rohrbach, Darrell, Canny, and Akata, in NIPS Interpretable ML Symposium, 2017

CAMEL: Causal Models to Explain Learning

Charles River Analytics (CRA), U. Mass, Brown

Explainable Model

Causal Modeling

- Experiment with the learned model (as a grey box) to learn an explainable, causal, probabilistic programming model

Explanation Interface

Narrative Generation

- Interactive visualization based on the generation of temporal, spatial narratives from the causal, probabilistic models

Challenge Problem

Autonomy

- Atari
- Starcraft

Data Analytics

- Pedestrian Detection (INRIA)
- Activity Recognition (ActivityNet)

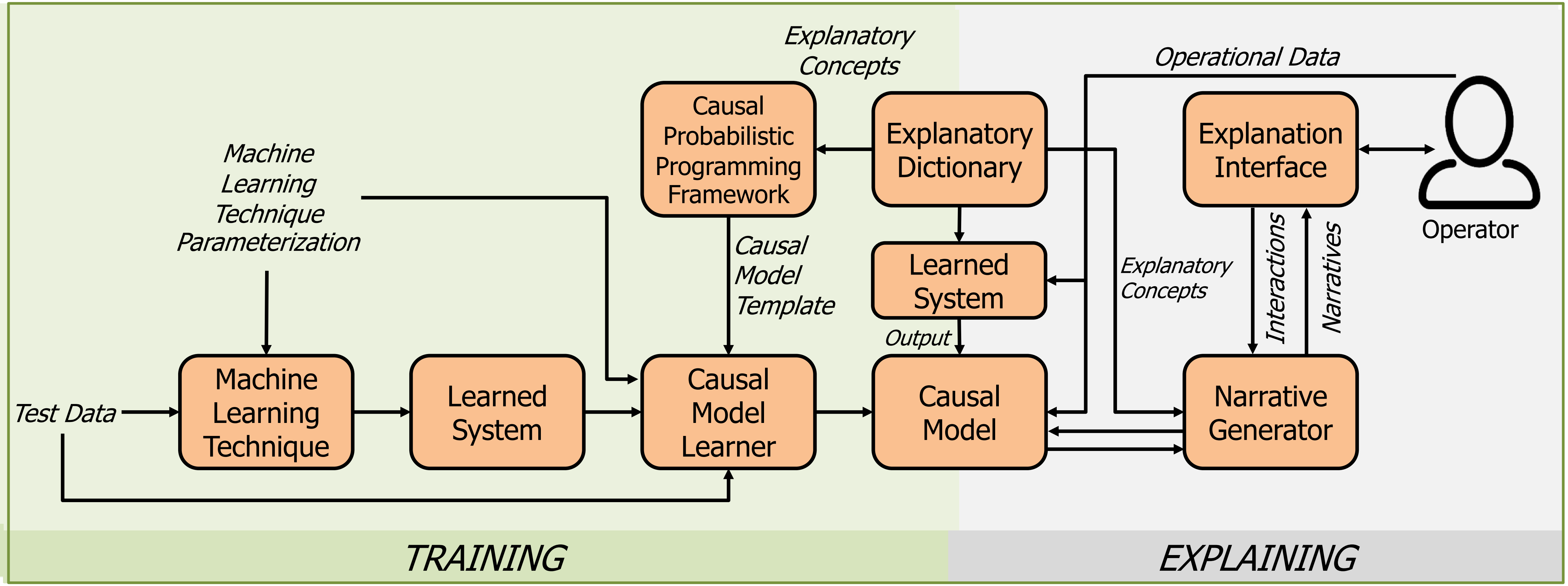
- **PI**: James Tittle (CRA)

- Jeff Druce (CRA)
- Avi Pfeffer (CRA)
- David Jensen (U. Mass)
- Michael Littman (Brown U.)
- James Niehaus (CRA)
- Emilie Roth (Roth Cognitive Engineering)
- Joe Gorman(CRA)
- James Tittle (CRA)

Causal Models to Explain Learning (CAMEL)

Charles River Analytics, U. Mass, Brown

Generate causal explanations of ML operation and present them to the user as intuitive narratives in an interactive, easy-to-use interface grounded in cognitive engineering theories



Learning and Communicating Explainable Representations

UCLA, Oregon State, Michigan State

Explainable Model

Pattern Theory+

Interpretable representations
- STC-AOG: spatial, temporal, and causal models
- STC-PG: scene and event interpretations in analytics
- STC-PG^{+}: task plans in autonomy

Theory of mind representations
- User's beliefs
- User's mental model of agent

Explanation Interface

3-Level Explanation
- Concept compositions
- Causal and counterfactual reasoning
- Utility explanations

Explanation representations:
- X-AOG: explanation model
- X-PG: explanatory parse graph as dialogue
- X-Utility: priority and loss for explanations

Challenge Problem

Autonomy
- Robot executing daily tasks in physics-realistic VR platform
- Autonomous vehicle driving (GTA5 game engine)

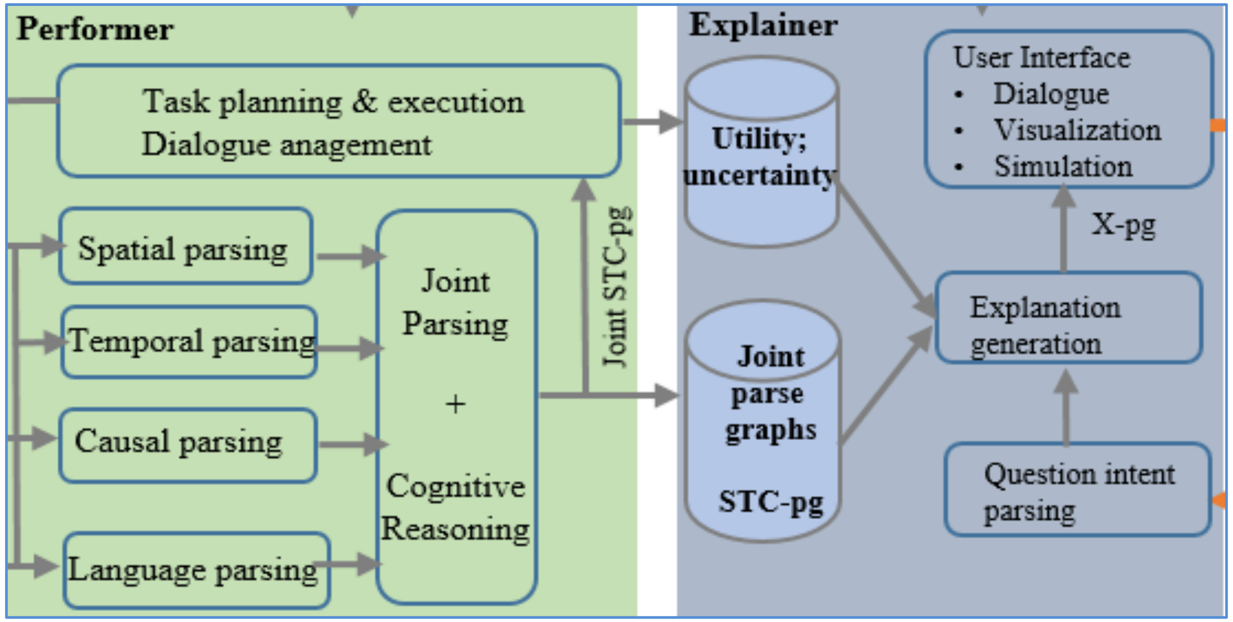
Data Analytics
- Network of video cameras for scene understanding and event analysis

- **PI**: Song-Chun Zhu (UCLA)

- Ying Nian Wu (UCLA)
- Sinisa Todorovic (OSU)
- Joyce Chai (Michigan State)

UCLA, Oregon State, Michigan State

System Architecture



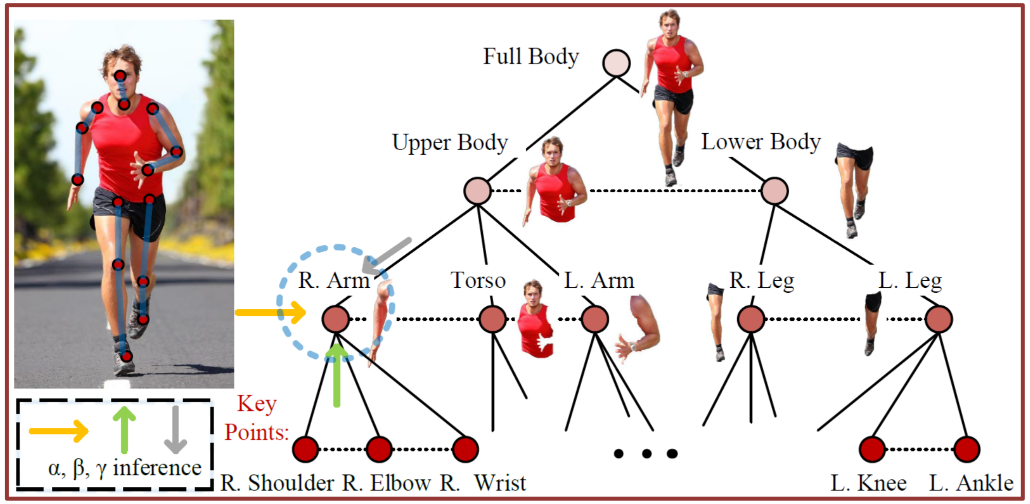
Performer

Outputs interpretable representations in a spatial, temporal, and causal parse graph (STC-PG)

Explainer

Outputs an explanatory parse graph (X-PG) in a dialogue process

STC Parse Graph



An attributed parse graph for a running person. Each node has 3 computing channels:

- α: grounding the node on DNN features;
- β: bottom-up;
- γ: top-down.

An explanation is represented as parse graph X-pg

xACT: Explanation-Informed Acceptance Testing of Deep Adaptive Programs

Oregon State University

Explainable Model

Adaptive Programs

- Explainable Deep Adaptive Programs (xDAPs) – a new combination of Adaptive Programs, Deep Learning, and explainability

Explanation Interface

Acceptance Testing

- Provides a visual and Natural Language explanation interface for acceptance testing by test pilots based on Information Foraging Theory

Challenge Problem

Autonomy

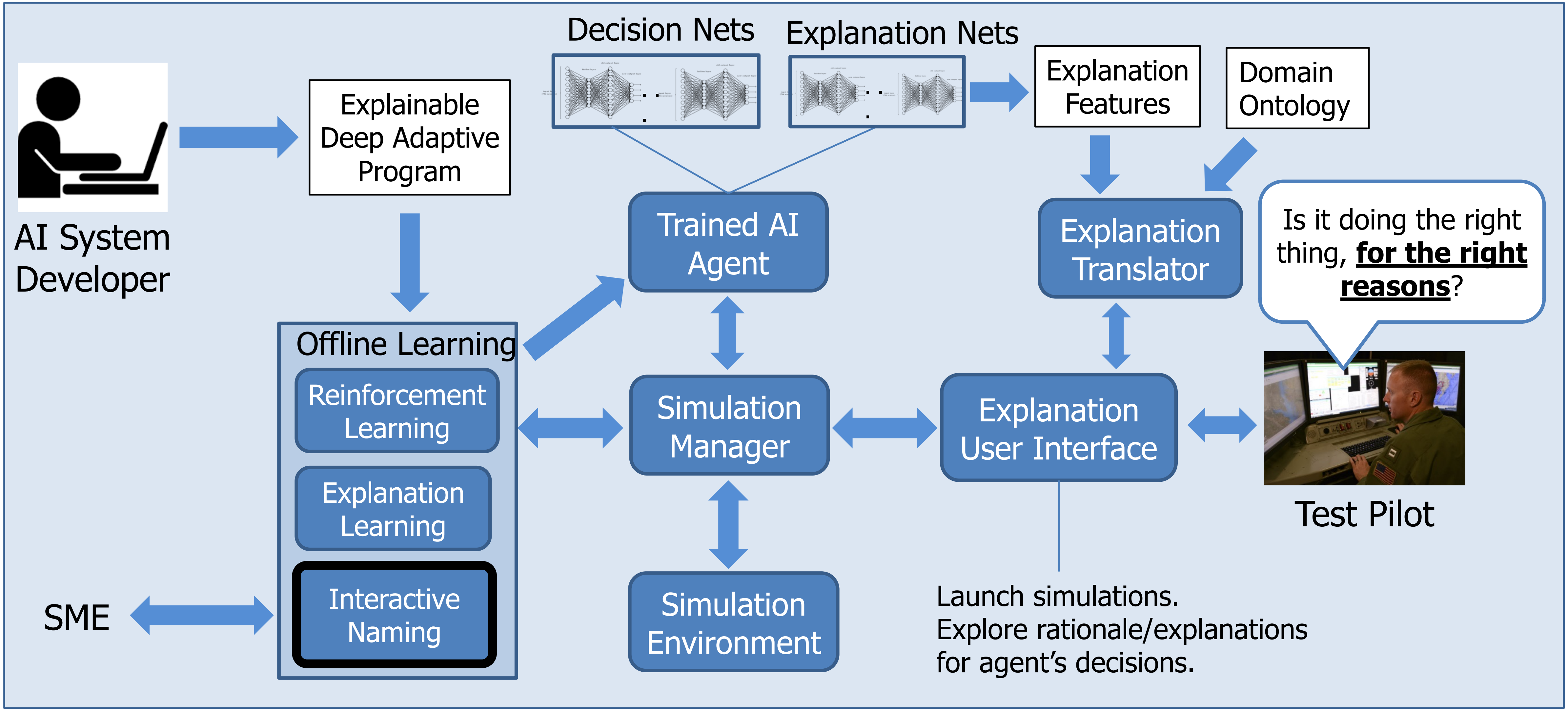
- Real-time Strategy Games based on custom designed game engine designed to support explanation
- Starcraft II

- **PI**: Alan Fern (OSU)

- Tom Dietterich (OSU)
- Fuxin Li (OSU)
- Prasad Tadepalli (OSU)
- Weng-Keen Wong (OSU)
- Margaret Burnett (OSU)
- Martin Erwig (OSU)
- Liang Huang (OSU)

xACT: Explanation-Informed Acceptance Testing of Deep Adaptive Programs

Oregon State University



COGLE: Common Ground Learning and Explanation

PARC, CMU, U. Edinburgh, U. Michigan, USMA, IHMC

Explainable Model	Explanation Interface	Challenge Problem
Cognitive Model 3-layer architecture • Learning Layer (DNNs) • Cognitive Layer (ACT-R Cognitive Model) • Explanation Layer (HCI)	**Interactive Training** • Interactive visualization of states, actions, policies, and values • Module for test pilots to refine and train the system	**Autonomy** • MAVSim wrapper over ArduPilot simulation environment • Value of Explanation framework for measuring explanation effectiveness

- **PI**: Mark Stefik (PARC)

- Honglak Lee (U. Michigan)
- Subramanian Ramamoorthy (U. Edinburgh)

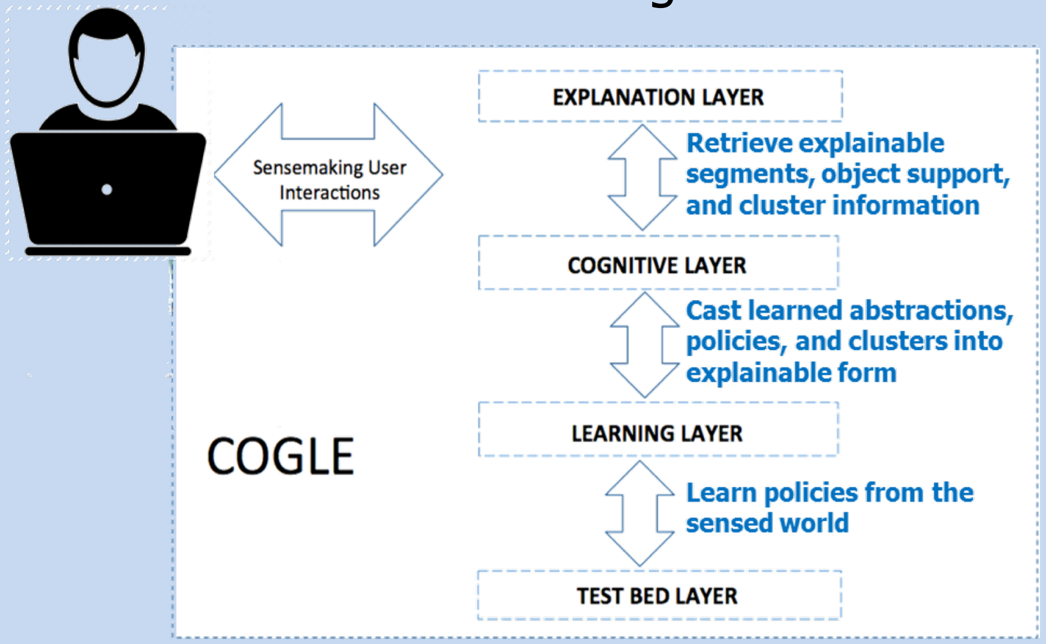
- Christian Lebiere (CMU)
- John Anderson (CMU)
- Robert Thomson (USMA)

- Michael Youngblood (PARC)

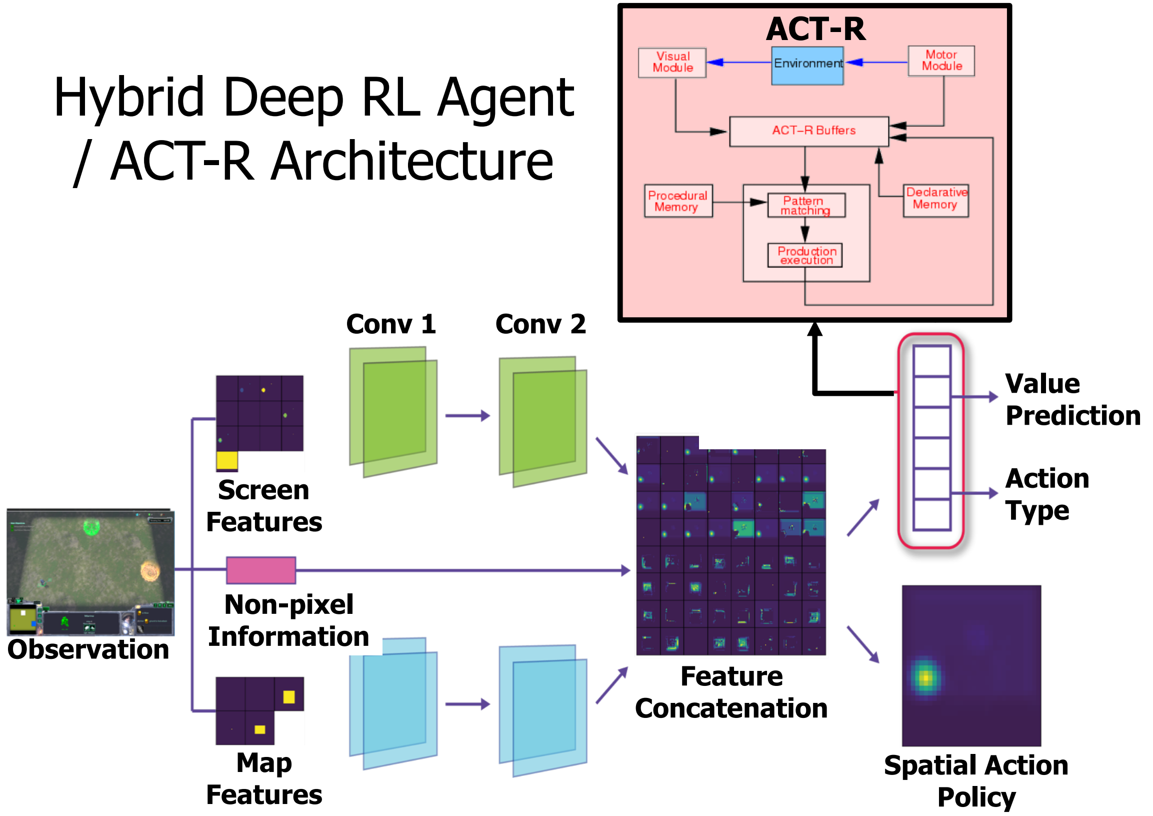
COGLE: Common Ground Learning and Explanation

PARC, CMU, U. Edinburgh, U. Michigan, USMA, IHMC

Layered Cognitive Architecture to Partition Explanation And Learning



Hybrid Deep RL Agent / ACT-R Architecture



XRL: Explainable Reinforcement Learning

Carnegie Mellon University

Explainable Model	Explanation Interface	Challenge Problem
Explainable RL (XRL)	**XRL Interaction**	**Autonomy**
• Create a new scientific discipline for Explainable Reinforcement Learning with work on new algorithms and representations	• Interactive explanations of dynamic systems • Human-machine interaction to improve performance	• Open AI Gym • Autonomy in the electrical grid • Mobile service robots • Self-improving educational software

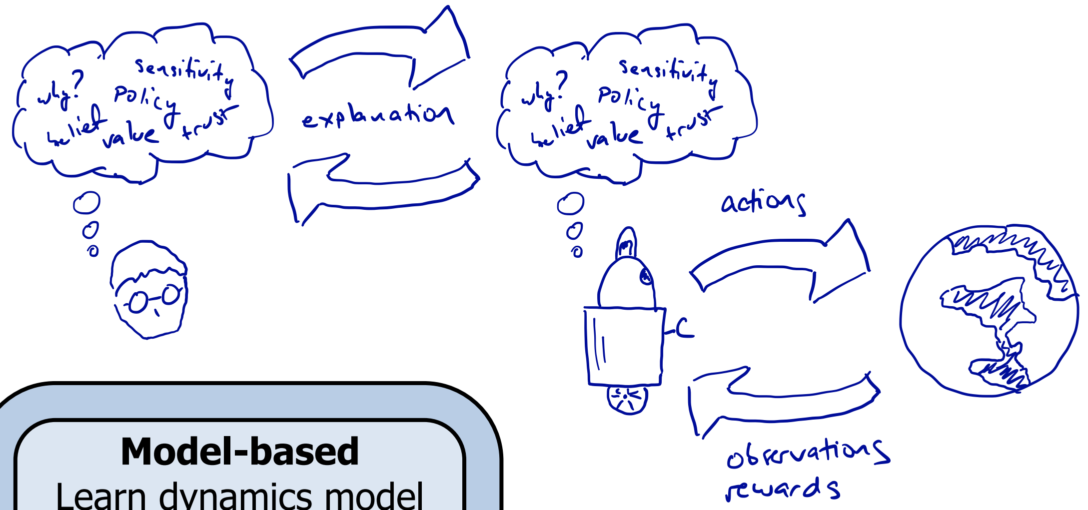
- **PI**: Zico Kolter (CMU)

- Geoff Gordon (CMU)
- Pradeep Ravikumar (CMU)

XRL: Explainable Reinforcement Learning

Carnegie Mellon University

Create a new discipline of explainable RL to enable dynamic human-machine interaction and adaptation for maximum team performance



Model-based
Learn dynamics model of environment, plan actions in model, and execute in real system

→ Improve model learning/ creation for RL agents to capture benefits of model-based approach

Model-free
Directly learn value and/or policy for the environment

→ For any type of RL approach, provide an explanation of why an agent acted in a certain way

Sensor inputs → Image segmentation → Rigid body simulation → iLQR-based MPC → Control

Differentiable Physics - Applies implicit differentiation to solutions of LCP to analytically derive a backpropagation update of next state with respect to previous state, control, and model parameters

DARE: Deep Attention-based Representations for Explanation

SRI International, U. Toronto, UCSD, U. Guelph

Explainable Model

Deep Learning

Multiple deep learning techniques
- Attention-based mechanisms
- Compositional NMNs
- GANs

Explanation Interface

Show & Tell Explanation

- DNN visualization
- Query evidence that explains DNN decisions
- Generate natural language justifications

Challenge Problem

Data Analytics

- VQA
  - Visual Gnome
  - Flickr30
- MovieQA

- **PIs**: Giedrius Burachas (SRI), Mohamed Amer (SRI)

- Xiao Lin (SRI)
- Ryan Villamil (SRI)
- Dejan Jovanovic (SRI)
- Avi Ziskind (SRI)
- Michael Wessel (SRI)

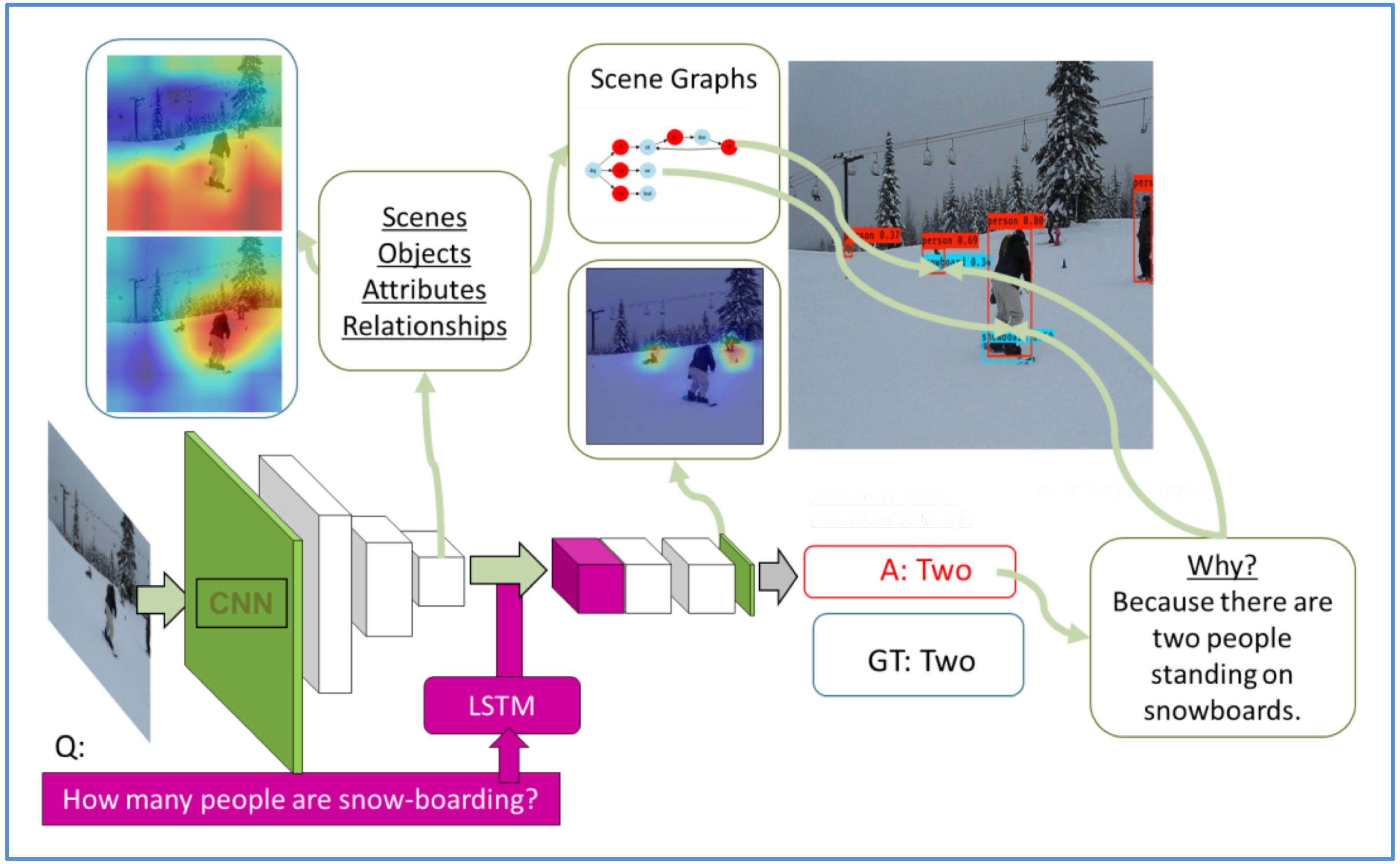
Richard R. Zemel (U. Toronto)
Sanja Fidler (U. Toronto)
David Duvenaud (U. Toronto)
Graham Taylor (U. Guelph)

- Jürgen Schulze (UCSD)

DARE: Deep Attention-based Representations for Explanation

SRI International, U. Toronto, UCSD, U. Guelph

Interpretable, Scene Graph-based VQA System with Active Attention



- Generate "show-and-tell" explanations with justifications of decisions accompanied by visualizations of input data used to generate inferences
- Scene and Situation Graphs, inferred from images and videos, support rich multimodal data analytics and explanations
- Scene Graphs guide attentional scanning for interpretable analytics

EQUAS: Explainable QUestion Answering System

Raytheon BBN, Georgia Tech, UT Austin, MIT

Explainable Model

Deep Learning

- Semantic labelling of DNN neurons
- DNN audit trail construction
- Gradient-weighted Class Activation Mapping

Explanation Interface

Argumentation Theory

- Comprehensive strategy based on argumentation theory
- NL generation
- DNN visualization

Challenge Problem

Data Analytics

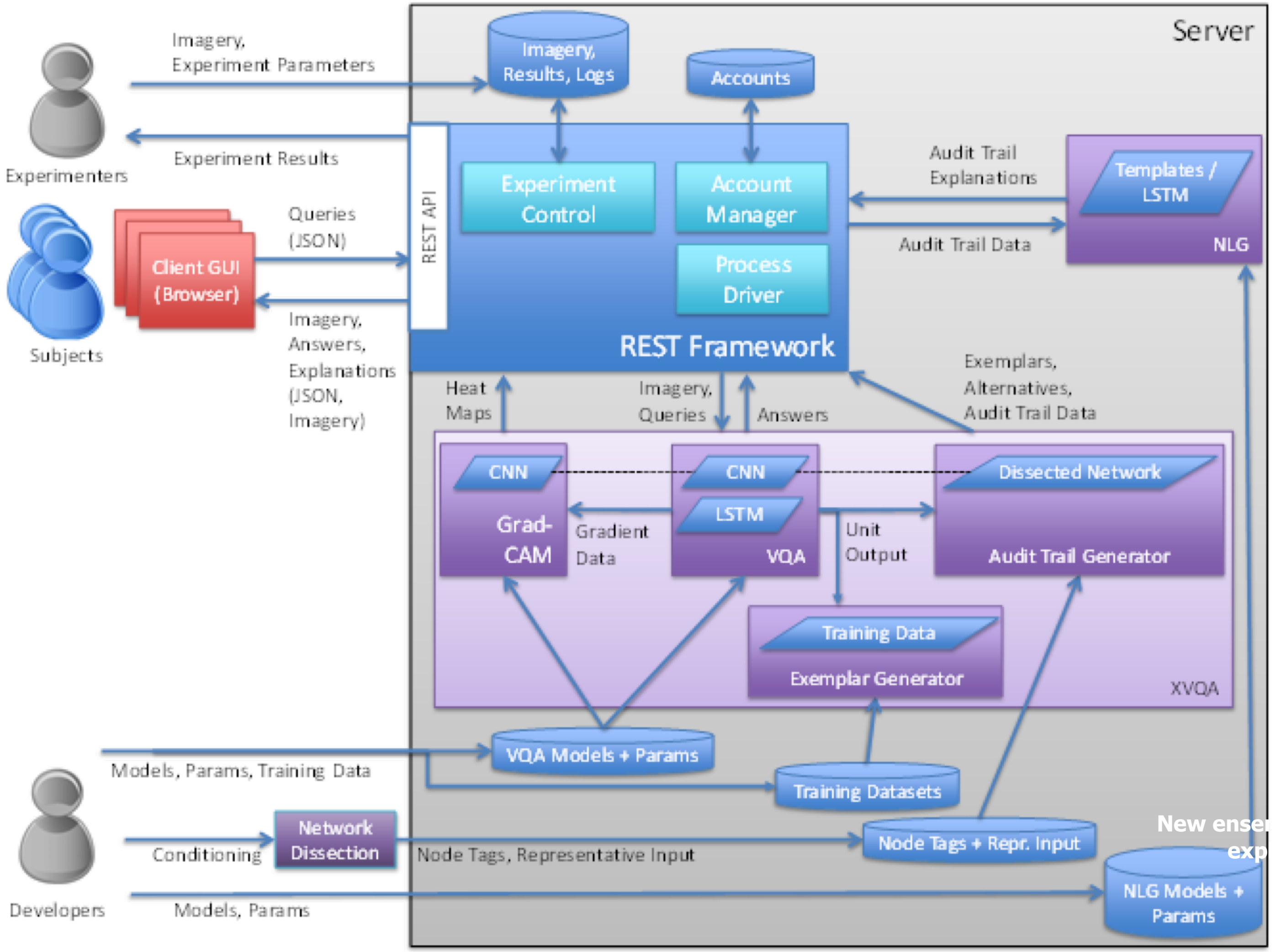
- VQA for images and video

- **PI**: William Ferguson (Raytheon BBN)

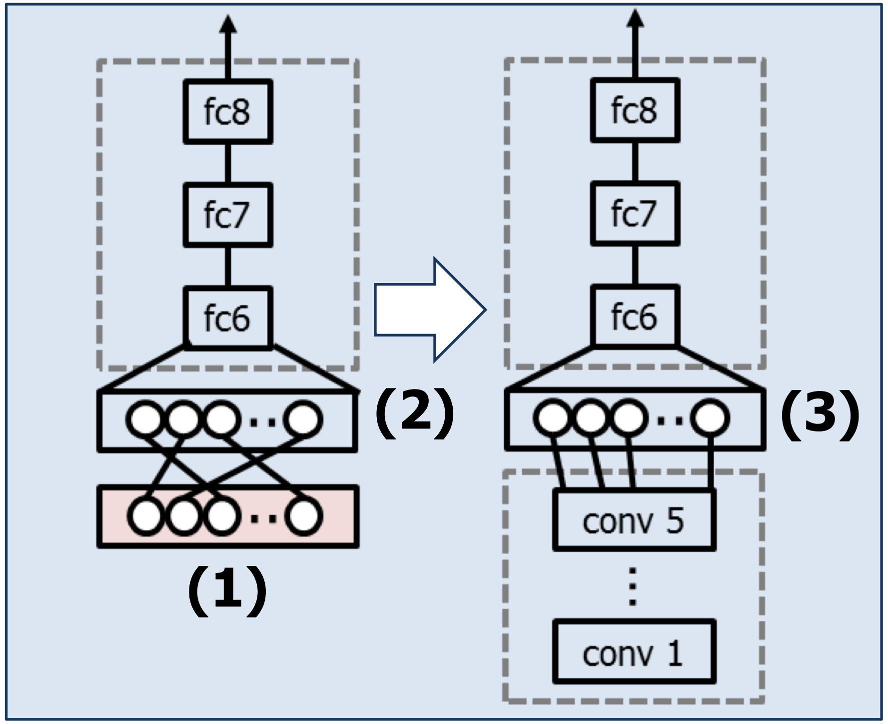
- Antonio Torralba (MIT)
- Ray Mooney (UT Austin)
- Devi Parikh (Georgia Tech)
- Dhruv Batra (Georgia Tech)

EQUAS: Explainable QUestion Answering System

Raytheon BBN, Georgia Tech, UT Austin, MIT



Improve the interpretability of units using a **new conditioning method** to retrain the network to intentionally include *concept detectors*



1) **Pick units from standard vocabulary**
2) **Train top part of net**
3) **Use top to train bottom**

Tractable Probabilistic Logic Models

UT Dallas, UCLA, Texas A&M, Indian Institute of Technology

Explainable Model

Probabilistic Logic

- Tractable Probabilistic Logic Models (TPLMs) – an important class of (non-deep learning) interpretable models

Explanation Interface

Decision Diagrams

- Enables users to explore and correct the underlying model as well as add background knowledge

Challenge Problem

Data Analytics

- Infer activities in multimodal data (video and text)
- Wetlab (biology) and TACoS (cooking) datasets

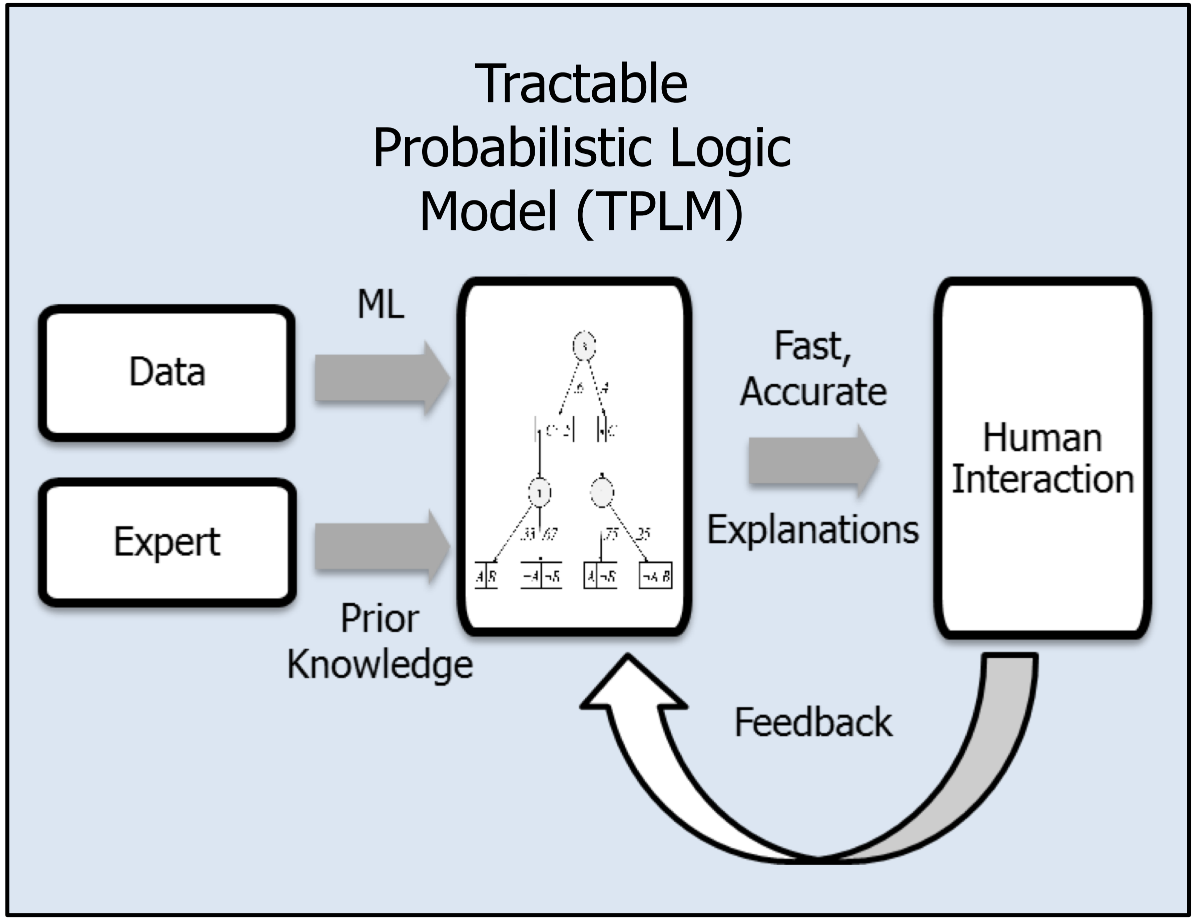
- **PI**: Vibhav Gogate (UT Dallas)

- Adnan Darwiche (UCLA)
- Guy Van Den Broeck (UCLA)
- Nicholas Ruozzi (UT Dallas)
- Eric Ragan (Texas A&M)
- Parag Singla (IIT-Delhi)

Tractable Probabilistic Logic Models

UT Dallas, UCLA, Texas A&M, Indian Institute of Technology

Use interpretable and tractable models based on well-founded principles from logic and probability theory



Find all videos in which a person peels oranges
(explanations are captions generated by TPLMs)

Person using his hands to peel oranges. I can see the orange skin

Person using his hands. I can see the orange skin and skinless orange

Person using his hands to peel oranges. I can see the orange skin on the table and peeled oranges

Transforming Deep Learning to Harness the Interpretability of Shallow Models

Texas A&M, Washington State

Explainable Model

Mimic Learning

- Mimic learning framework combines DL models for prediction and shallow models for explanations
- Interpretable learning algorithms extract knowledge from DNNs for relevant explanations

Explanation Interface

Interactive Visualization

- Interactive visualization over multiple views, using heat maps and topic modeling clusters to show predictive features

Challenge Problem

Data Analytics

- Multiple tasks using data from Twitter, Facebook, ImageNet, and news websites

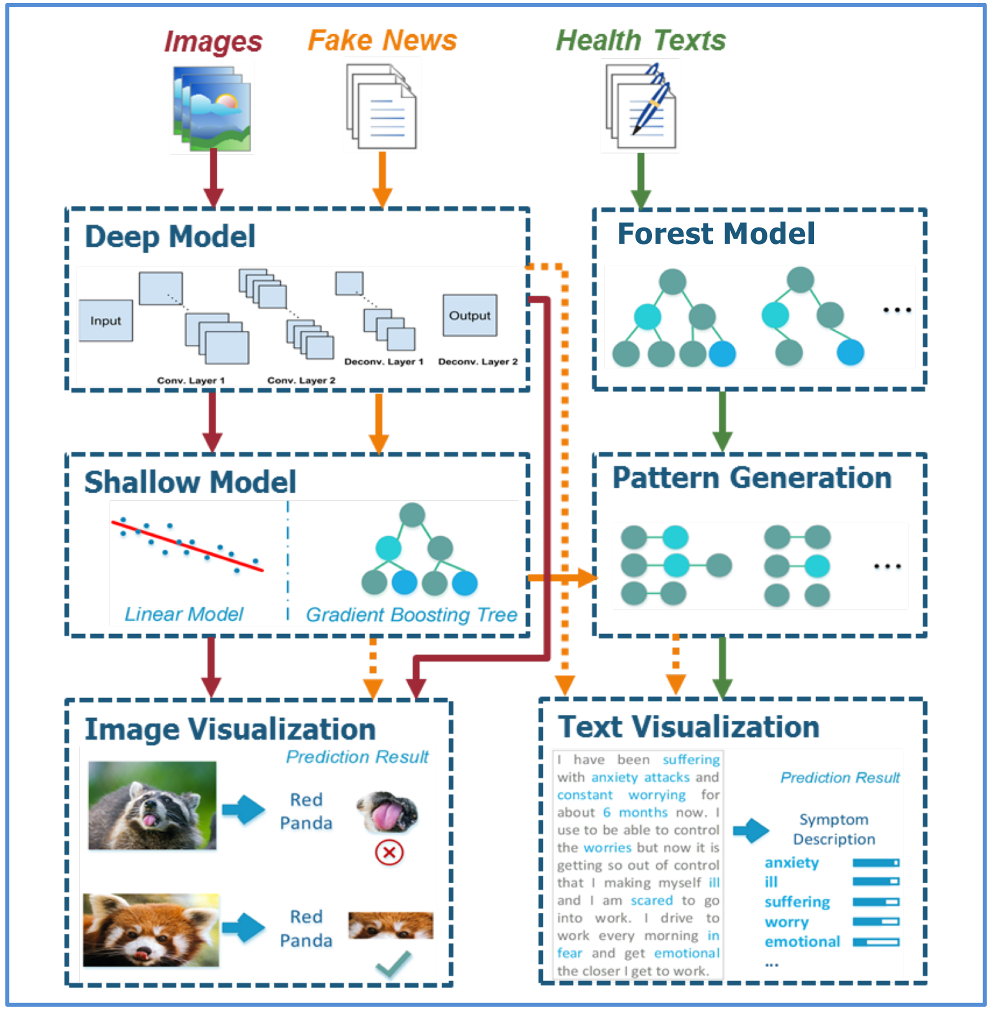
- **PI**: Xia Hu (Texas A&M)

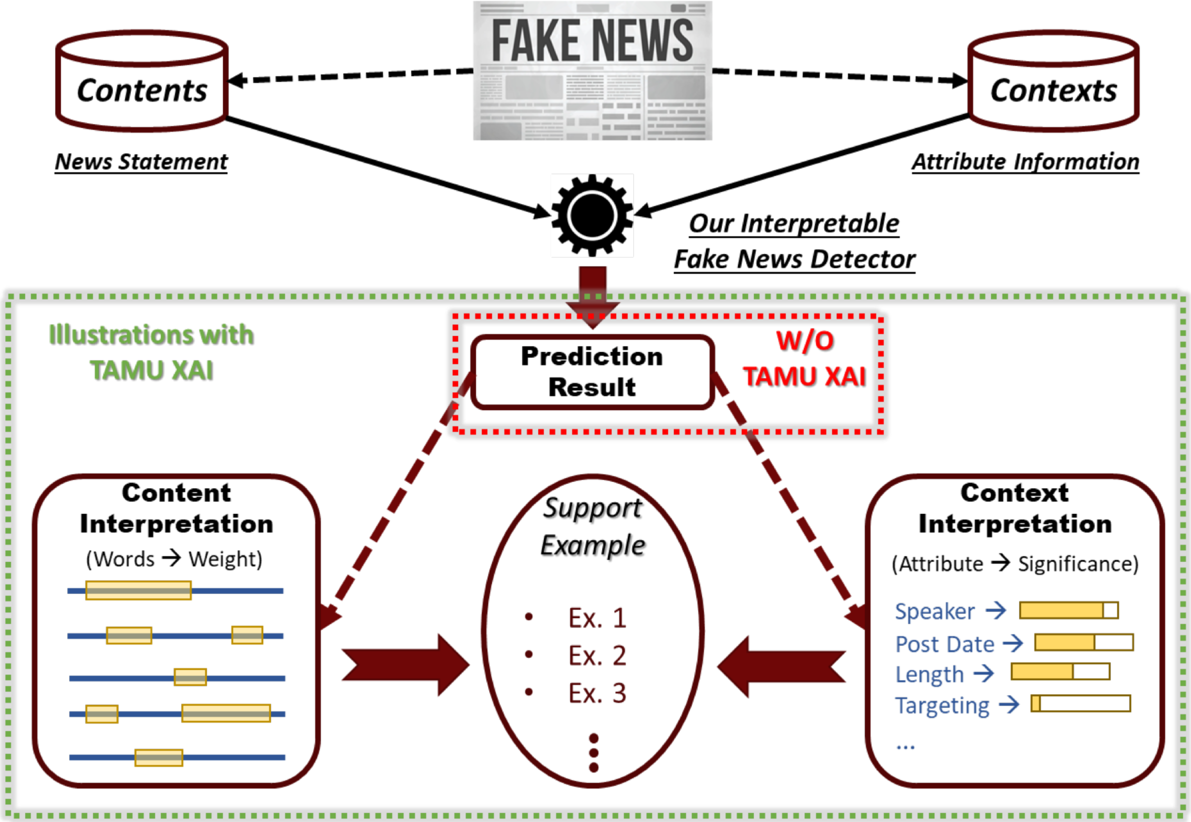
- Shuiwang Ji (Wash. State)
- Eric Ragan (Texas A&M)

Transforming Deep Learning to Harness the Interpretability of Shallow Models

Texas A&M, Washington State

Develop an end-to-end interpretable deep learning infrastructure with image and text datasets





Model Explanation by Optimal Selection of Teaching Examples

Rutgers University

Explainable Model

Model Induction

- Select the optimal training examples to explain model decisions based on Bayesian Teaching

Explanation Interface

Bayesian Teaching

- Example-based explanation of
  - Full model
  - User-selected sub-structure
  - User submitted examples

Challenge Problem

Data Analytics

- Image processing
- Text corpora
- VQA
- Movie events

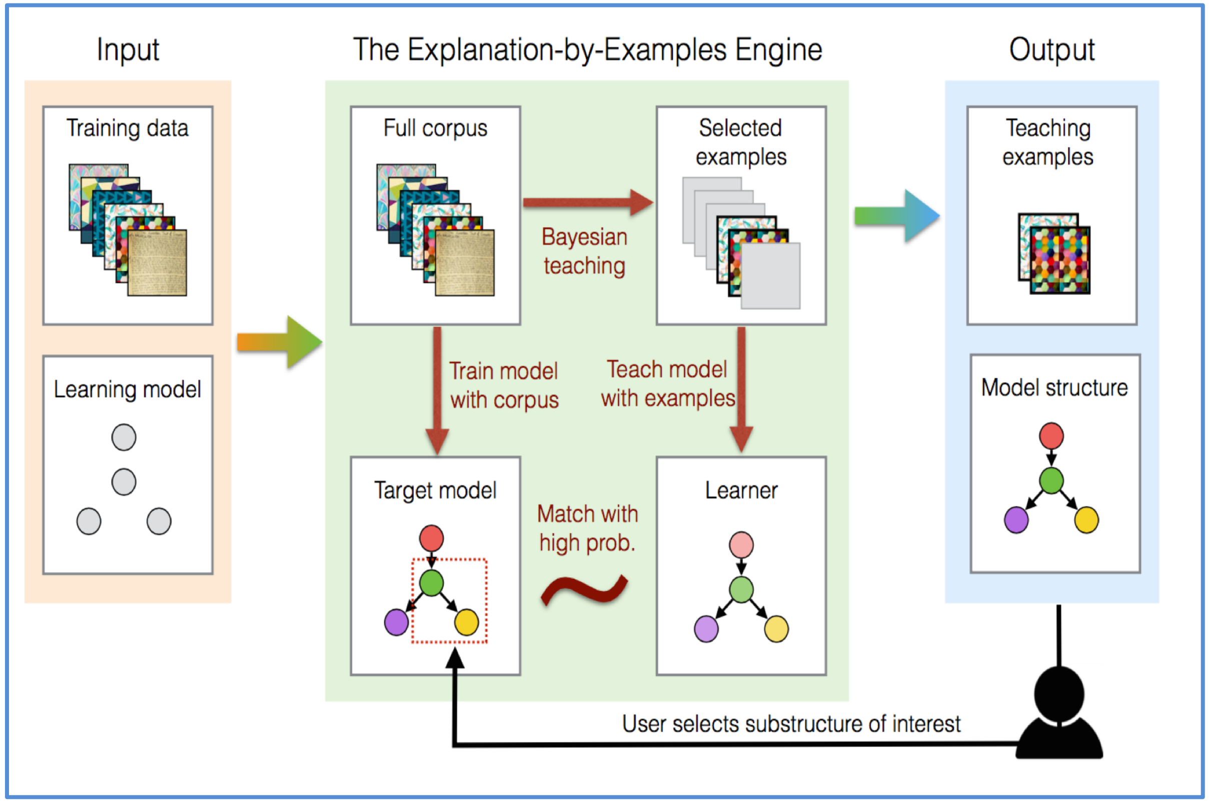
- **PI**: Patrick Shafto (Rutgers)

- Scott Cheng-Hsin Yang (Rutgers)

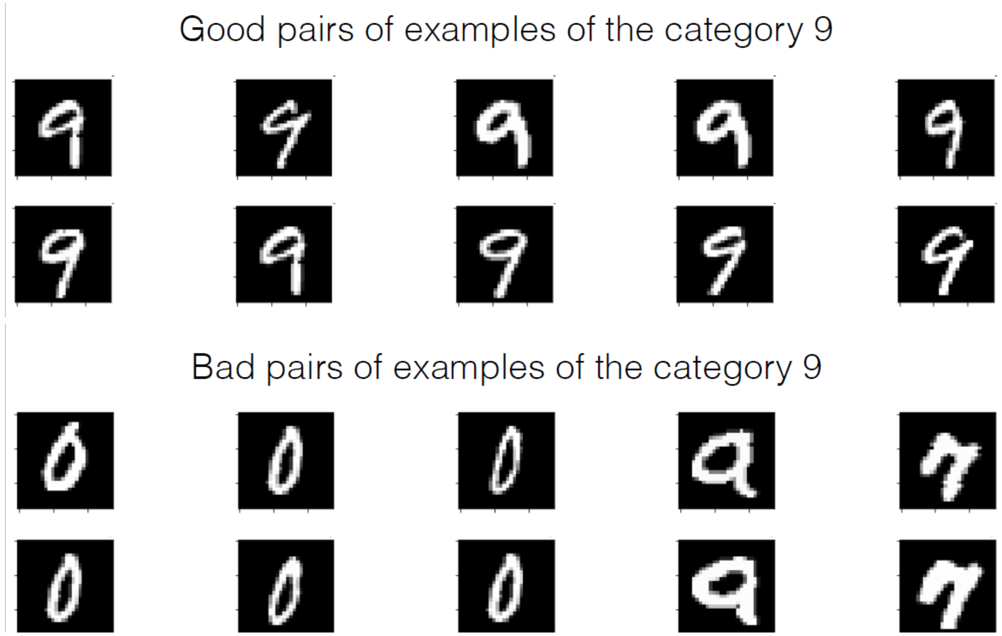
Model Explanation by Optimal Selection of Examples

Rutgers University

Extend Bayesian teaching to enable automatic explanation by selecting the subset of data that are most representative of the model's generative process

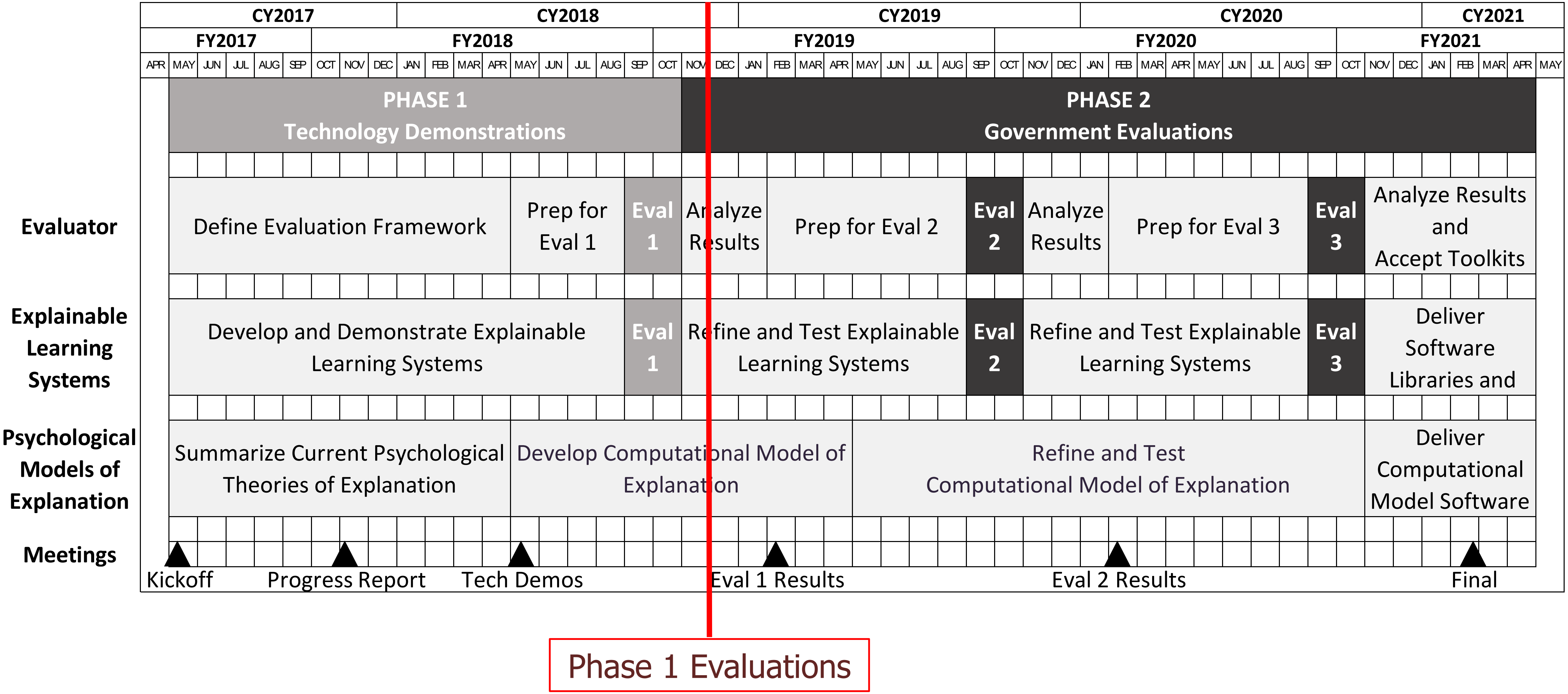


Good and bad examples for teaching a category (illustrates model strengths and weaknesses)



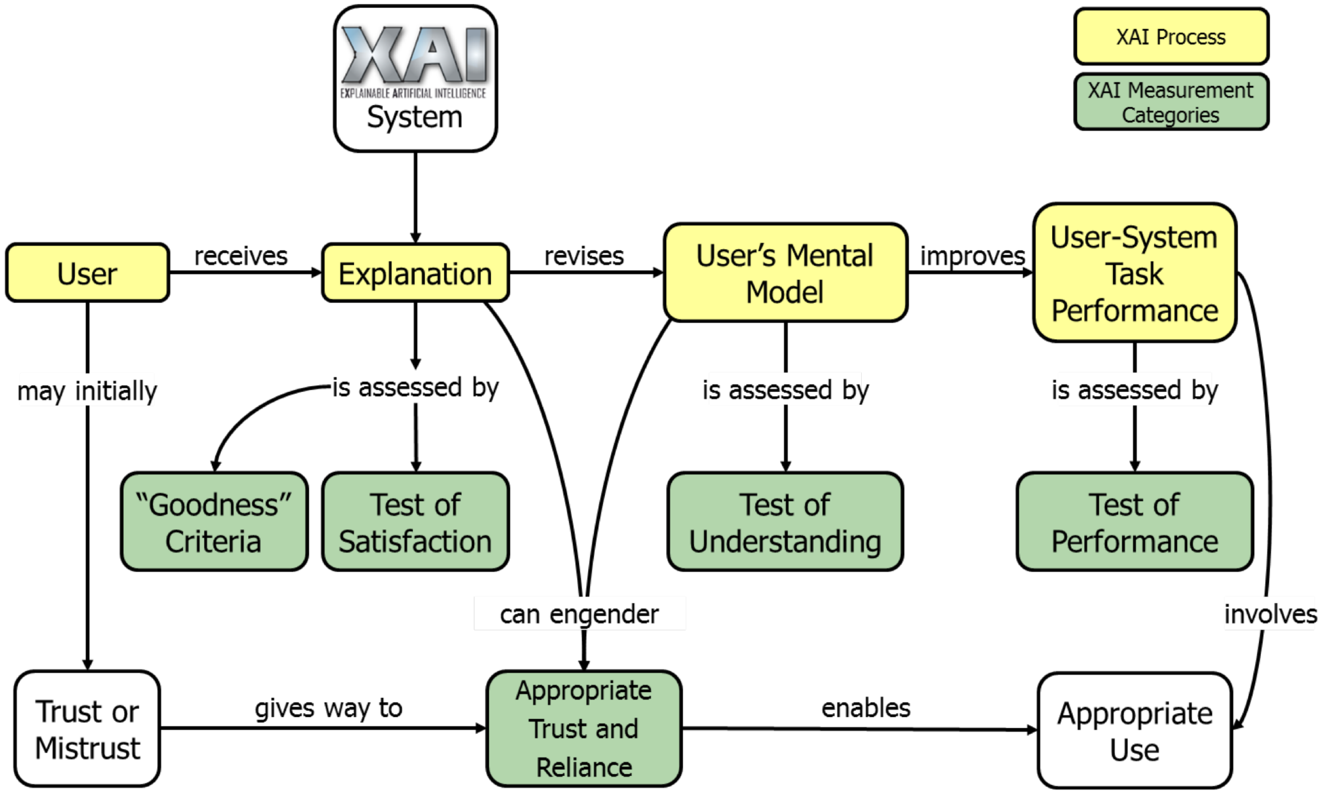
XAI Program Schedule

XAI
EXPLAINABLE ARTIFICIAL INTELLIGENCE



Evaluation Framework

Explanation Process & Measures



Experimental Conditions

Without Explanation - The explainable learning system is used to perform a task without providing an explanation to the user

With Explanation - The explainable learning system is used to perform a task and generates explanations for every recommendation or decision it makes, and every action it takes

Partial Explanation - The explainable learning system is used to perform a task and generates only partial or ablated explanations (to assess various explanation features)

Control - A baseline state-of-the-art non-explainable system is used to perform a task

DARPA

www.darpa.mil